Sheer Curation for Experimental Data and Provenance

Tobias Blanke[a] and Mark Hedges[a]
[a]Centre for e-Research, King's College London, UK

Background: Sheer Curation

This poster presents an approach to the curation of the experimental data and processes of a group of scientific researchers in the field of biophysics. Recent research has demonstrated that carrying out digital curation and preservation activities in the early stages of data creation is more cost-effective compared to the potential loss that can be incurred through the destruction of data, for example because of the need to recreate the data, or the loss of an organisation's reputation. On the one hand, decisions taken during the early stages of a digital object's lifecycle frequently have consequences for the preservation strategies that can be applied at a later date; on the other hand, if digital objects are being preserved so that they can be reused in an informed manner, account has to be taken of the different practices of researchers across disciplines and the different characteristics of the data they create or gather. One approach to integrating research processes into data management has been termed sheer curation. Here, digital curation activities are integrated into the workflow of the researchers creating or capturing data. The word 'sheer' is used to express the 'lightweight and virtually transparent' way in which these curation processes are integrated, with minimal disruption to researchers' normal working practices.

References

Borgman, C. L. (2007), Scholarship in the digital age: Information, infrastructure, and the Internet (Cambridge, MA: MIT Press).

Brase, J. (2009), Datacite - a global registration agency for research data, COINFO '09: Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 257-261.

Moreau, L., Groth, P. (2010), Open Provenance Model (OPM) XML Schema Specification. Latest version http://openprovenance.org/model/opmx-20101012.

Moreau, L., Cliord, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., and Van den Bussche, J. (2011), The Open Provenance Model core specification (v1.1). Future Generation Computer Systems, Vol. 27, No. 6, 743-756.

Pepe, A., Mayernik, M., Borgman, C.L., Van de Sompel, H. (2009), From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web, Journal of the American Society for Information Science and Technology, Vol. 61, No. 3, 567-582.

Rumsey, A. S. (ed.) (2010), Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information, Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

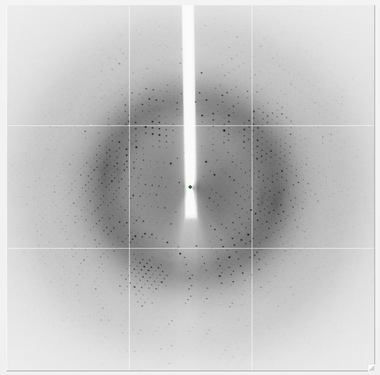
Simmhan, Y., Plale, B., Gannon, D. (2005), A survey of data provenance in e-science, SIGMOD Record, Vol. 34, No. 3, 31-36.

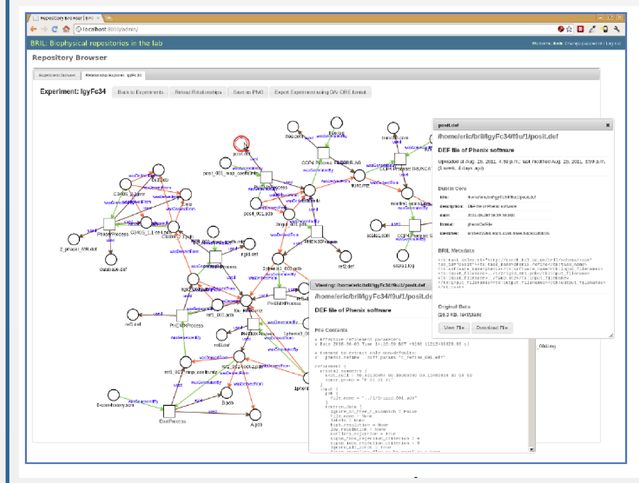
Zhao, J., Goble, C., Stevens, R., Turi, D. (2008), Mining Taverna's Semantic Web of Provenance, Concurrency and Computation: Practice and Experience, Vol. 20, No. 5, 463-472.

Zhao, J. (2010), Open Provenance Model Vocabulary Specification. Latest version: http://purl.org/net/opmv/ns-20101006.

Case Study 1: Macromolecular crystallography

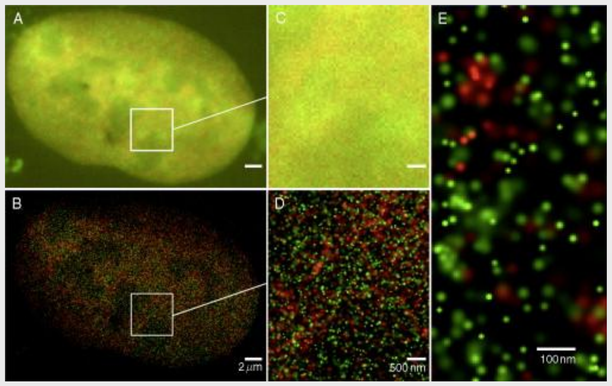
Macromolecular crystallography addresses the determination of the structure of large molecules, such as proteins, using X-ray diffraction In high-level terms, an X-ray beam is directed at a crystal of the substance under investigation from many angles, resulting in a set (typically 360, although sometimes more) of diffraction images. Each image contains several hundred spots, whose location and intensity are determined, using specialised software, and then combined to produce a model of the atomic co-ordinates of the protein. In this process a large number of files is created but only a small number of the resulting files are published and kept.





Case Study 2: Biological Nanoimaging

Biological nanoimaging involves the use of microscopes to capture high-resolution images of biological samples, on the one hand to carry out research into cell and tissue structures, and on the other to develop new methods of and algorithms for digital imaging and processing. The data include pseudo-3D representations. The same datasets may be processed many times using different image analysis techniques, and many raw images are processed when developing new analysis tools. Again, much of the information generated in this process is not currently curated or retained.



Implementation Results

Managing and organising digital research assets

We embedded digital library services within the experimental workflows of the researchers, which involve the execution of a variety of interactive tools to process and analyse the raw data on their local desktop computers. The implementation used a lightweight client, running on the researcher's computer, to 'scavenge' information from the researcher's work area and transfer it to the digital library environment, where it is interpreted and stored. This made it possible to capture automatically domain-specific metadata and other contextual information that is only available when research is undertaken. This approach can also be made to work in situations where it is not possible to use the monitoring client on the researcher's desktop. To implement this, the directory is submitted for deposit in its entirety at the end of the experiment.

Managing entire experimental workflows

In the researcher's local work space on their desktop computer, the story of the experiment is represented implicitly in a variety of information such as the location of files in a directory hierarchy. In the digital library environment, experimental workflows were modelled explicitly as compound objects incorporating data, metadata and provenance information in order to verify published results or reproduce the processing and data. We used the Open Provenance Model (OPM) as an emerging standard for modelling provenance that aims to enable the digital representation of the provenance of any object. To publish the experimental processes, we have used a linked data model, with URIs resolving to a representation of the digital objects that are connected by links described using the OPM vocabulary, and potentially other ontologies.

Conclusion and Future Work

Our work has shown the advantages of a sheer curation approach for involving research processes in the management of data, not just for preservation but also for publication purposes. However, the approach does require a quite detailed understanding of the researchers' work practices, which has to be elicited from them via user engagement activities that can be very time-consuming, particularly when the development team has no background knowledge of the discipline in question. The extent to which this effort is justified would have to be addressed on a case-by-case basis.